دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

کارشناسی ارشد علوم کامپیوتر گرایش داده کاوی

پروژه شماره دو درس داده کاوی

نگارش

حدیث حق‌شناس جزی

استاد راهنما

مهدی قطعی

استاد مشاور

بهنام یوسفی مهر

مهر ۱۴۰۱

چکیده

در این پروژه دیتاست املاک و مستغلات ملبورن مورد بررسی قرار گرفته است. بررسی ها به کمک روش های کاهش ابعاد و رسم انواع نمودار های مختلف, بر روی انواع ستون های ویژگی دیتا انجام شده است. همانطور که در ادامه خواهیم دید, میان قیمت انواع املاک در ملبورن نظیر ویلا, خانه, سوییت و ... , فاصله از مرکز شهر , متراژ ملک و نوع ملک و ... روابطی وجود دارد که مفصلا به آنها خواهیم داد.

فصل اول

مقدمه

در این پروژه سعی شده‌است که یک دیتاست مناسب برای مصور سازی انتخاب شود. سپس به معرفی دیتاست, علت انتخاب آن و انجام تحلیل های آماری داده های مربوطه و مصورسازی آن‌ها می‌پردازیم.

همچنین برای درک بهتر دیتاست انواع اندازه‌های مختلف بر روی ویژگی‌های متفاوت آزمایش شده است و در آخر به تفسیر این نتایج پرداخته می‌شود.

فصل دوم

۲-۱ دیتاست Melbourne housing snapshot

دیتاست املاک و مستغلات ملبورن, در ستون های مجزا به انواع اطلاعات این املاک پرداخته است و به دلیل رونق این املاک از مجموعه داده ایجاد شده توسط تونی پینو جمع آوری شده است, به این امید که به سوالاتی نظیر سوالات زیر پاسخ دهد:

آیا می توان با تحلیل این دیتاست, ترند بعدی انواع املاک را در ملبورن پیش بینی کرد؟ و یا حداقل میتوان به کمک داده ها به تخمین حدودی قیمت مناسب برای ملک مورد نظرمان رسید؟ و یا میشود امکان وجود داشتن خانه ای با ویژگی های مورد نیازمان را تخمین بزنیم؟

مجموعه ویژگی ها(ستون ها) در این دیتاست شامل آدرس، نوع املاک(سوییت، ویلا، خانه و ...) ، تعداد پارکینگ ، روش فروش، تعداد اتاق ها، قیمت ملک ، نماینده فروش، تاریخ فروش، فاصله از مرکز شهر, مکان جغرافیایی ، قدمت و متراژ می‌باشد.

۲-۲ بررسی دیتاست

در ابتدا پس از فراخوانی داده، ۵ سطر اول را نمایش میدهیم. این داده شامل ۱۳۵۸۰ سطر و ۲۱ ستون میباشد.

```
In [3]: #project2 /data mining

        import pandas as pd
        #seaborn, a Python graphing library
        import seaborn as sns
        import matplotlib.pyplot as plt

        df = pd.read_csv("Desktop/melb_data.csv") # the melbourne house dataset is now a Pandas DataFrame

        df.head()
```

Out[3]:

	Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Distance	Postcode	...	Bathroom	Car	Landsize	BuildingArea	YearBuilt	Cou
0	Abbotsford	85 Turner St	2	h	1480000.0	S	Biggin	3/12/2016	2.5	3067.0	...	1.0	1.0	202.0	NaN	NaN	
1	Abbotsford	25 Bloomburg St	2	h	1035000.0	S	Biggin	4/02/2016	2.5	3067.0	...	1.0	0.0	156.0	79.0	1900.0	
2	Abbotsford	5 Charles St	3	h	1465000.0	SP	Biggin	4/03/2017	2.5	3067.0	...	2.0	0.0	134.0	150.0	1900.0	
3	Abbotsford	40 Federation La	3	h	850000.0	PI	Biggin	4/03/2017	2.5	3067.0	...	2.0	1.0	94.0	NaN	NaN	
4	Abbotsford	55a Park St	4	h	1600000.0	VB	Nelson	4/06/2016	2.5	3067.0	...	1.0	2.0	120.0	142.0	2014.0	

5 rows × 21 columns

اطلاعات کلی داده به شرح زیر است:

```
In [4]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13580 entries, 0 to 13579
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Suburb         13580 non-null  object
 1   Address        13580 non-null  object
 2   Rooms          13580 non-null  int64
 3   Type           13580 non-null  object
 4   Price          13580 non-null  float64
 5   Method         13580 non-null  object
 6   SellerG        13580 non-null  object
 7   Date           13580 non-null  object
 8   Distance       13580 non-null  float64
 9   Postcode       13580 non-null  float64
 10  Bedroom2       13580 non-null  float64
 11  Bathroom       13580 non-null  float64
 12  Car            13518 non-null  float64
 13  Landsize       13580 non-null  float64
 14  BuildingArea   7130 non-null   float64
 15  YearBuilt      8205 non-null   float64
 16  CouncilArea    12211 non-null  object
 17  Lattitude      13580 non-null  float64
 18  Longtitude     13580 non-null  float64
 19  Regionname     13580 non-null  object
 20  Propertycount  13580 non-null  float64
dtypes: float64(12), int64(1), object(8)
memory usage: 2.2+ MB
```

همانطور که مشاهده میشود در این دیتاست به طور کلی ۳ نوع داده عددی، اعشاری و شی است

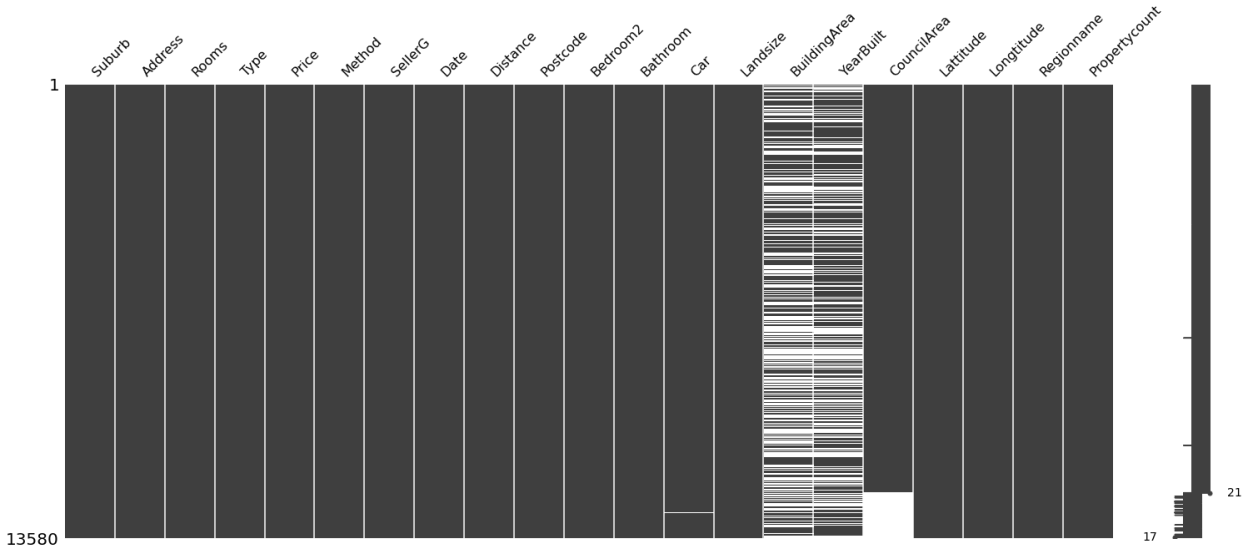
۲-۲-۱ اندازه های آماری:

```
: df.describe()
```

:

	Rooms	Price	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuilt
count	13580.000000	1.358000e+04	13580.000000	13580.000000	13580.000000	13580.000000	13518.000000	13580.000000	7130.000000	8205.000000
mean	2.937997	1.075684e+06	10.137776	3105.301915	2.914728	1.534242	1.610075	558.416127	151.967650	1964.684217
std	0.955748	6.393107e+05	5.868725	90.676964	0.965921	0.691712	0.962634	3990.669241	541.014538	37.273762
min	1.000000	8.500000e+04	0.000000	3000.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1196.000000
25%	2.000000	6.500000e+05	6.100000	3044.000000	2.000000	1.000000	1.000000	177.000000	93.000000	1940.000000
50%	3.000000	9.030000e+05	9.200000	3084.000000	3.000000	1.000000	2.000000	440.000000	126.000000	1970.000000
75%	3.000000	1.330000e+06	13.000000	3148.000000	3.000000	2.000000	2.000000	651.000000	174.000000	1999.000000
max	10.000000	9.000000e+06	48.100000	3977.000000	20.000000	8.000000	10.000000	433014.000000	44515.000000	2018.000000

در جدول بالا تمام موارد میانگین، مینیمم، چارک اول تا سوم، ماکسیمم و تعداد داده ها(عددی) مشخص شده اند. در ستون آخر تعداد داده ها ۸۲۰۵ عدد میباشد. این به معنای وجود تعداد زیادی اطلاعات گمشده یا میسینگ دیتاست. به همین دلیل با استفاده از کتابخانه msno به پیگیری این موضوع خواهیم پرداخت.



همانطور که حدس زدیم در ستون های قدمت ملک و موقعیت مکانی ملک تعداد زیادی داده موجود نمیباشد.

به همین دلیل این ستون را (قدمت ملک) پس از کشیدن چند نمودار(نسبت به خود ستون، مانند نمودار جعبه ای) از داده ها (به علاوه ستون آدرس به دلیل بی مصرف بودن در تحلیل ها) به صورت زیر حذف خواهیم کرد:

```
df.drop(['BuildingArea','Address'],axis=1).head(5)
```

Out[7]:

	Suburb	Rooms	Type	Price	Method	SellerG	Date	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	YearBuilt	CouncilArea	Lattitude
0	Abbotsford	2	h	1480000.0	S	Biggin	3/12/2016	2.5	3067.0	2.0	1.0	1.0	202.0	NaN	Yarra	-37.7996
1	Abbotsford	2	h	1035000.0	S	Biggin	4/02/2016	2.5	3067.0	2.0	1.0	0.0	156.0	1900.0	Yarra	-37.8079
2	Abbotsford	3	h	1465000.0	SP	Biggin	4/03/2017	2.5	3067.0	3.0	2.0	0.0	134.0	1900.0	Yarra	-37.8093
3	Abbotsford	3	h	850000.0	PI	Biggin	4/03/2017	2.5	3067.0	3.0	2.0	1.0	94.0	NaN	Yarra	-37.7969
4	Abbotsford	4	h	1600000.0	VB	Nelson	4/06/2016	2.5	3067.0	3.0	1.0	2.0	120.0	2014.0	Yarra	-37.8072

۳-۲ بررسی ستون ها

ده عدد از ویژگی ها (تعداد اتاق، نوع ملک، قیمت، وضعیت ملک ، املاکی مرتبط، فاصله از شهر و ...) را به کمک تعداد انواع دسته ها یا فراوانی آنها میتوان تحلیل کرد. شمایه کلی به صورت زیر است :

```
In [38]: import numpy as np
         #gosaste
         ft= pd.DataFrame(df, columns =['Rooms','Type','Price','Method',
                                        'SellerG','Distance','YearBuilt',
                                        'Regionname','Bathroom', 'Car'])
         ft
```

Out[38]:

	Rooms	Type	Price	Method	SellerG	Distance	YearBuilt	Regionname	Bathroom	Car
0	2	h	1480000.0	S	Biggin	2.5	NaN	Northern Metropolitan	1.0	1.0
1	2	h	1035000.0	S	Biggin	2.5	1900.0	Northern Metropolitan	1.0	0.0
2	3	h	1465000.0	SP	Biggin	2.5	1900.0	Northern Metropolitan	2.0	0.0
3	3	h	850000.0	PI	Biggin	2.5	NaN	Northern Metropolitan	2.0	1.0
4	4	h	1600000.0	VB	Nelson	2.5	2014.0	Northern Metropolitan	1.0	2.0
...	...	...	...	...	...	...	...	...	...	...
13575	4	h	1245000.0	S	Barry	16.7	1981.0	South-Eastern Metropolitan	2.0	2.0
13576	3	h	1031000.0	SP	Williams	6.8	1995.0	Western Metropolitan	2.0	2.0
13577	3	h	1170000.0	S	Raine	6.8	1997.0	Western Metropolitan	2.0	4.0
13578	4	h	2500000.0	PI	Sweeney	6.8	1920.0	Western Metropolitan	1.0	5.0
13579	4	h	1285000.0	SP	Village	6.3	1920.0	Western Metropolitan	1.0	1.0

13580 rows × 10 columns

فراوانی انواع ملک:

```
In [41]: ft.Method.value_counts()

Out[41]: S     9022
         SP    1703
         PI    1564
         VB    1199
         SA      92
         Name: Method, dtype: int64
```

تعداد ۹۰۲۲ ملک از ۱۳۵۸۰ ملک به فروش رفته است. دومین جایگاه را املاک بدون خریداری دارند که قبلا به فروش رفته است.

فراوانی انواع پارکینگ :

```
In [47]: ft.Car.value_counts()
Out[47]: 2.0     5591
         1.0     5509
         0.0     1026
         3.0      748
         4.0      506
         5.0       63
         6.0       54
         8.0        9
         7.0        8
         10.0       3
         9.0        1
         Name: Car, dtype: int64
```

اکثر املاک شامل ۲ پارکینگ میشنود . پس از آن املاک ۱ پارکینگه در صدر جدول هستند. ۷ درصد آنها بدون پارکنیگ و ۵ درصد آنها بیش از ۲ پارکینگ دارند.

فراوانی انواع سرویس بهداشتی:

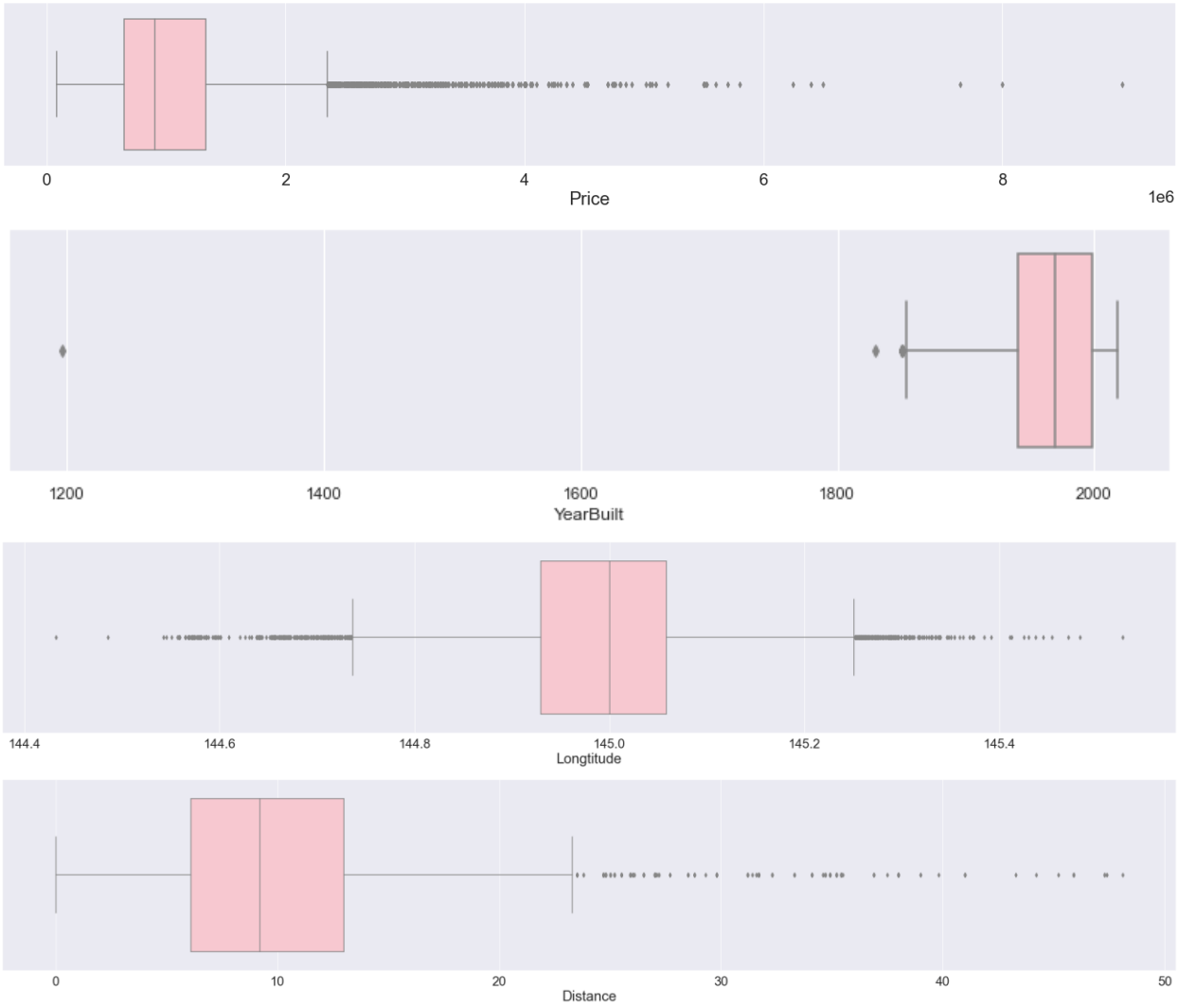
```
In [46]: ft.Bathroom.value_counts()
Out[46]: 1.0    7512
         2.0    4974
         3.0     917
         4.0     106
         0.0      34
         5.0      28
         6.0       5
         8.0       2
         7.0       2
         Name: Bathroom, dtype: int64
```

حدود ۵۵ درصد املاک یک سرویس بهداشتی دارند. و حدود ۳۶ درصد دو سرویس بهداشتی دارند. با توجه به انواع متراژ و اینکه اکثریت املاک در ملبورن ویلایی هستند (در ادامه خواهیم دید)، میتوان گفت که تعداد سرویس بهداشتی بالا مورد استقبال سازندگان املاک ملبورن نمی‌باشد.

فصل سوم

مصورسازی و تفسیر

۲-۲ نمودار جعبه ای:



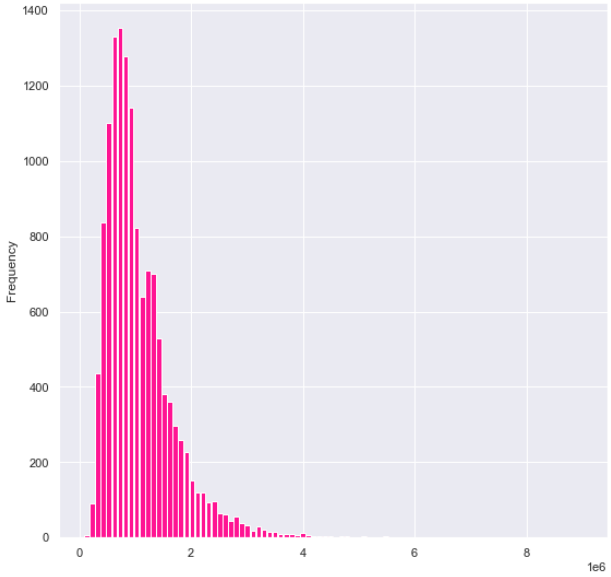
نمودار جعبه ای ستون های قیمت، قدمت، طول ملک، فاصله از مرکز شهر آورده شده است. همانطور که دیده میشود داده های قیمت نشانگر حدود قیمت نرمال اما شامل داده های پرت میباشند. این داده های پرت در قیمت املاک در نمودار های دیگر نیز قابل مشاهده است. با اینکه تعداد آنها به نسبت کلی داده ها کم است اما قابل توجه است.

در نمودار قدمت مشاهده میکنیم که قدمت اکثر املاک در بازه سالهای ۱۹۵۰ تا ۲۰۰۰ است.

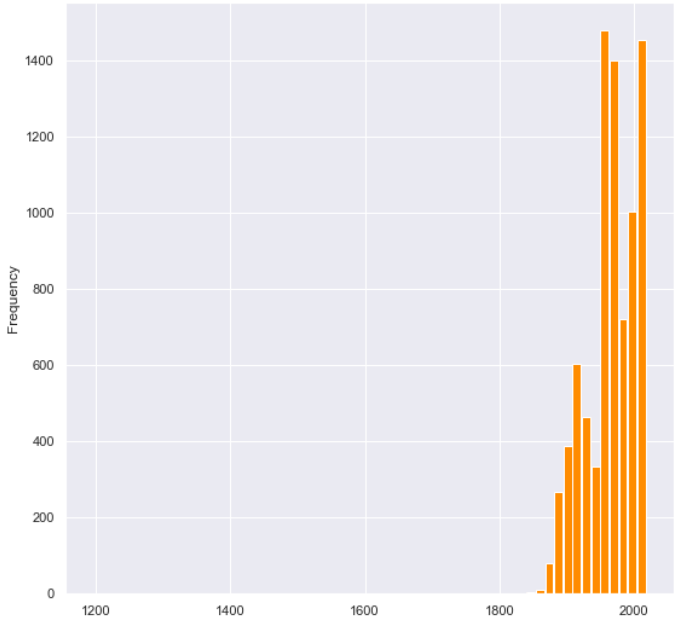
در نمودار طول ملک به داده های پرت زیادی داریم اما متوسط متراژ در بازه ۱۴۵ متر ثابت است.

در نمودار فاصله اکثریت املاک ۶ الی ۱۳ مایل است. تعدادی از املاک فواصل بسیار زیادتری نیز دارند اما به دلیل پایین بودن داده های پرت اینطور به نظر میرسد که داده های جمع آوری شده در سطح شهر و در اطراف مرکز شهر هستند و خیلی از مرکز شهر شعاع دورتری از داده ها موجود نیست.

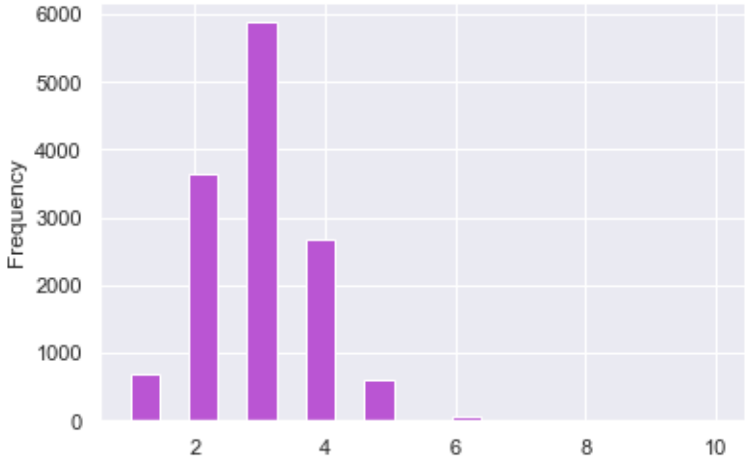
۲-۳ نمودار هیستوگرام:



توزیع قیمت املاک



توزیع قدمت املاک. (پس از این نمودار، ستون قدمت حذف خواهد شد. با توجه به ۲کوهانه بودن نمودار و شیب های تند آن میتوان پی برد که داده های گم شده در این ستون داده های مورد نیاز و خوبی را از دسترس خارج کرده اند و برای تحلیل های بهتر، بهتر است که این ستون را از میان تحلیل های خود حذف کنیم)



توزیع تعداد اتاق به طور کلی بیشترین فراوانی اتاق شامل اتاق های ۳ تخته و سپس ۲ تخته میشود. با وجود متراژهای متفاوت تعداد اتاق های ۵ و ۶ تخته به نسبت بسیار کم میباشد.

۳-۳ نمودار همبستگی:

نمودار همبستگی بین داده های عددی به صورت زیر است.

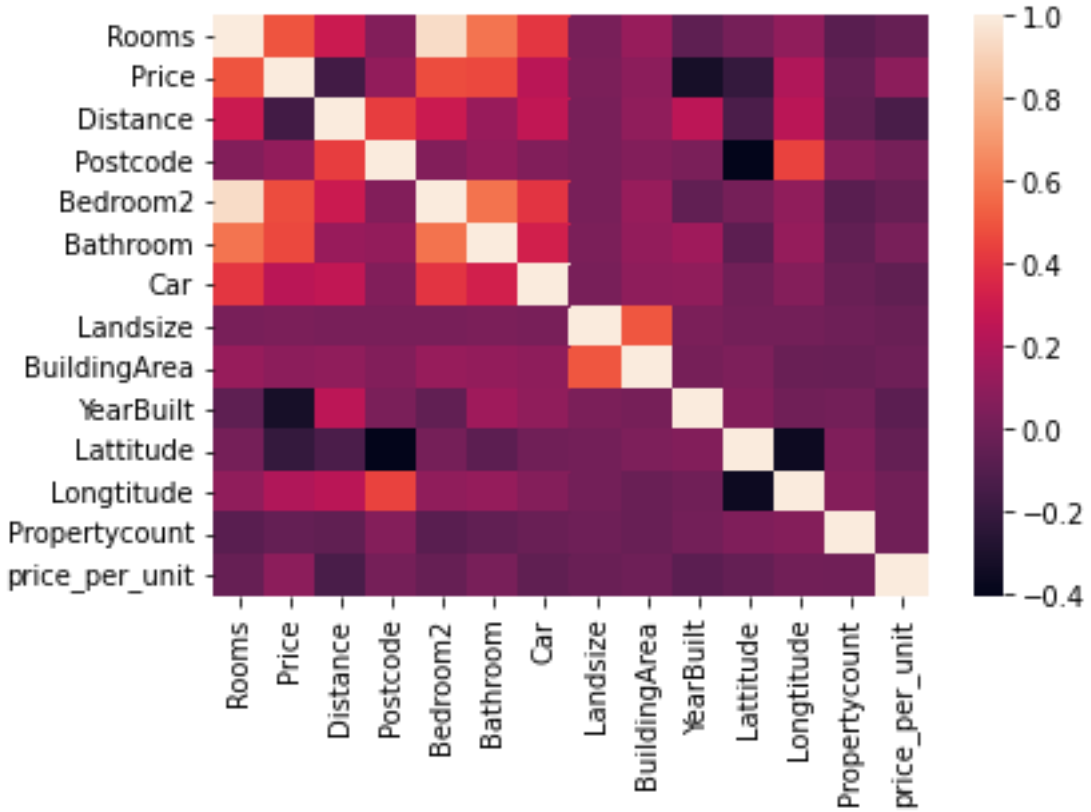
طبق این نمودار میان رنگ های مشکی و سفید همسبتگی منفی و مثبت وجود دارد. تنها همبستگی معنادار مابین ستون ها به شرح زیر میباشد:

قدمت و قیمت همبستگی منفی دارند. هرچه قدمت ساختمان بیشتر شود قیمت آن کاهش پیدا میکند.

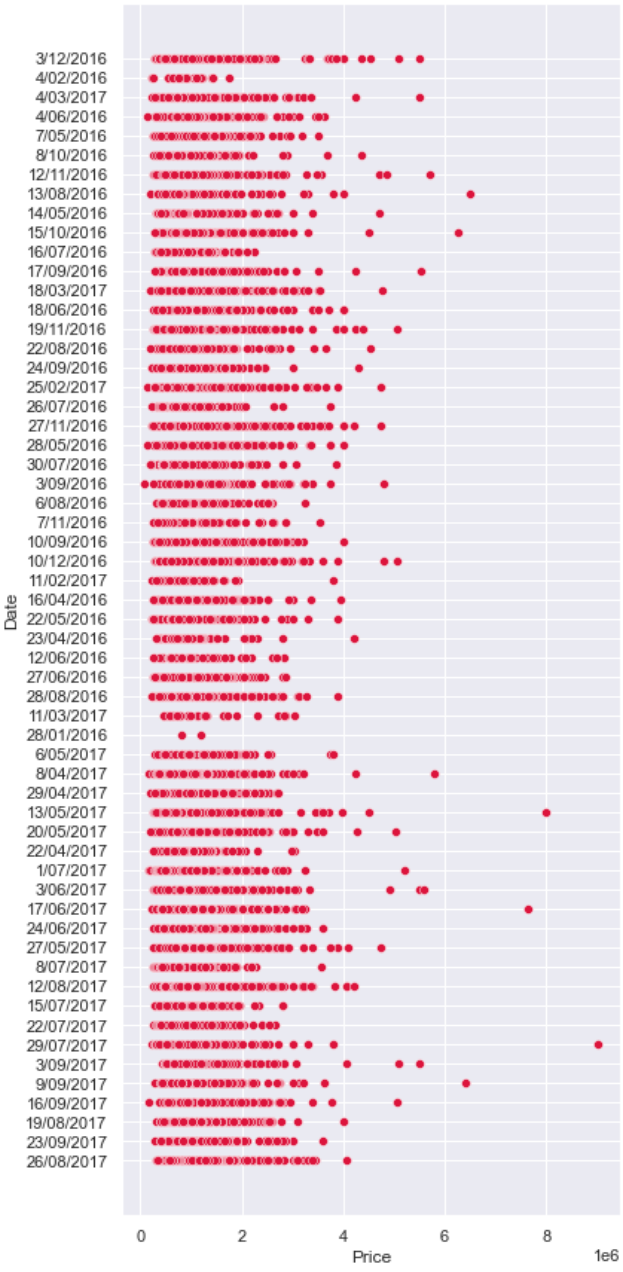
قیمت و فاصله از مرکز شهر همبستگی منفی دارند. هرچه فاصله از مرکز شهر کمتر شود قیمت افزایش پیدا میکند.

هر ستون با خودش همبستگی مثبت دارد!(واضح است)

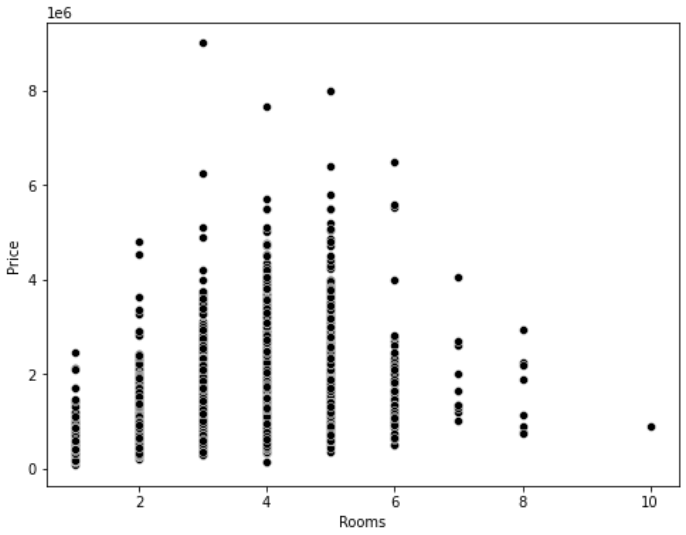
تعداد اتاق و تعداد اتاق خواب همبستگی مثبت دارد(واضح است(تعداد اتاق= تعداد اتاق خواب+۱))



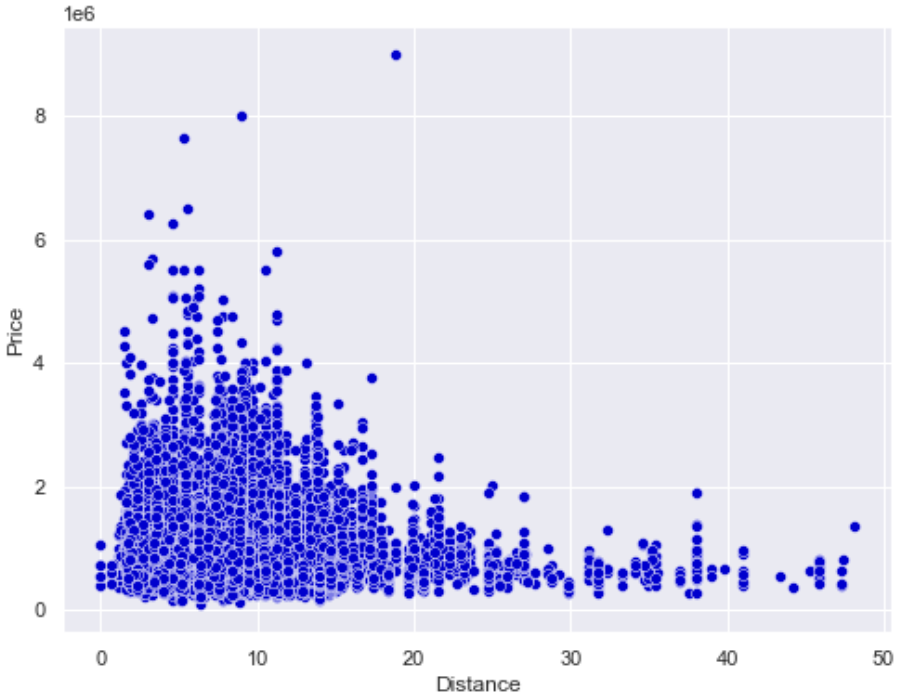
۳-۴ نمودار نقطه ای:



با توجه به نمودار بالا رابطه خاصی میان قیمت املاک و گذشت زمان نمیباشد.(اما تعداد داده های پرت(قیمت های پرت)کاهش داشته است)

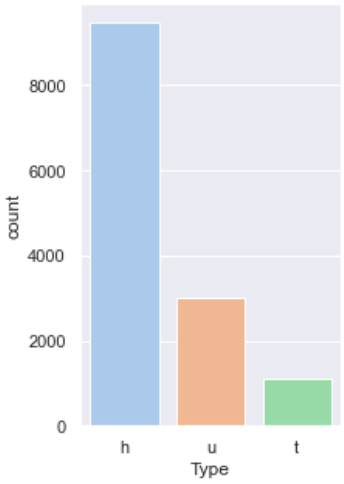


در این نمودار به میزان افزایش تعداد اتاق افزایش نسبی قیمت خانه راه مشاهده میکنیم.(همچنین داده پرتی داریم که پایین ترین قیمت با بالاترین تعداد اتاق را شامل میشود)

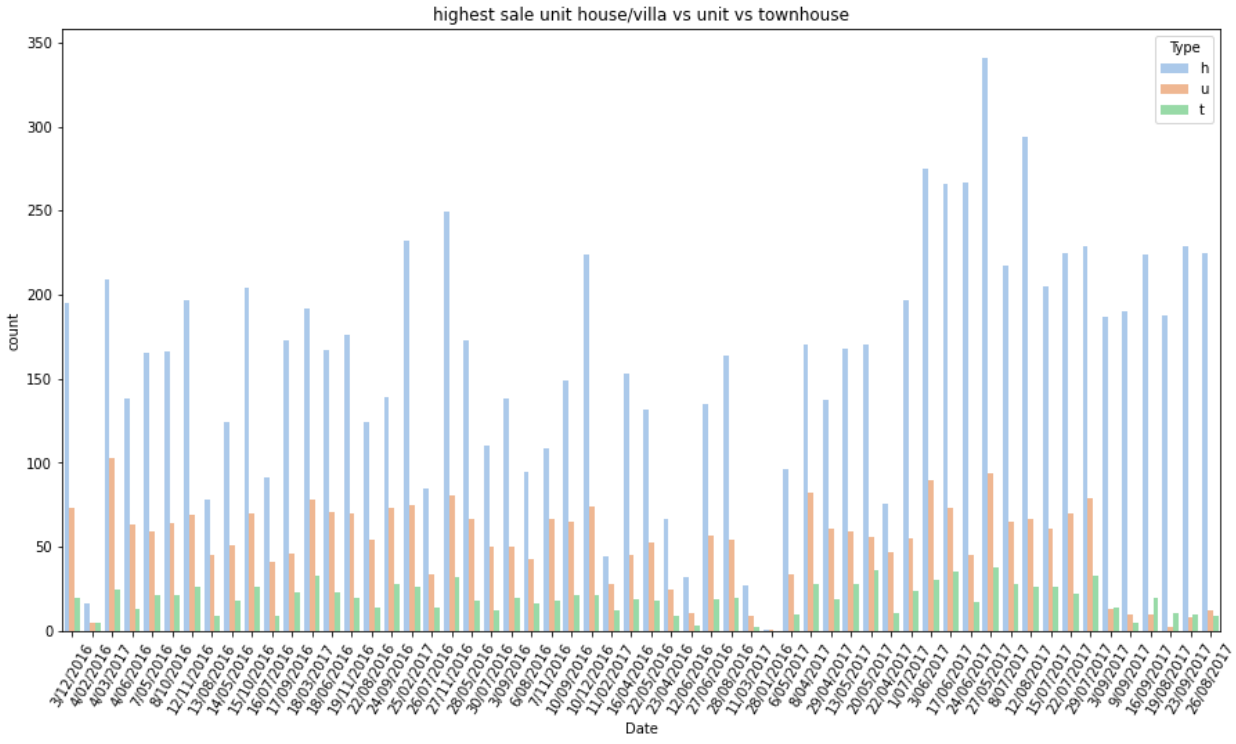


نمودار بالا نشانگر رابطه مستقیم میان فاصله از مرکز شهر و قیمت ملک میباشد. لذا هرچه فاصله از مرکز شهر کمتر باشد قیمت خانه ها به میزان چشم گیری بالاتر میرود. همچنین در این نمودار فراوانی زیاد خانه های نزدیک به مرکز شهر را نیز میبینیم.

۳-۵ سایر نمودار ها:



فراوانی انواع املاک با اختلاف زیادی به سمت املاک ویلایی و یا خانه های بیرون شهری و سپس واحد ها و سوییت ها میباشد.



در اینجا میزان فروش املاک در طول سال های ۲۰۱۶ تا ۲۰۱۷ را داریم. مدل املاک با رنگ های آبی و نارنجی و سبز جدا شده است. همانطور که مشاهده میشود همواره فروش واحد ها و سوئیت ها از خانه های ویلایی و

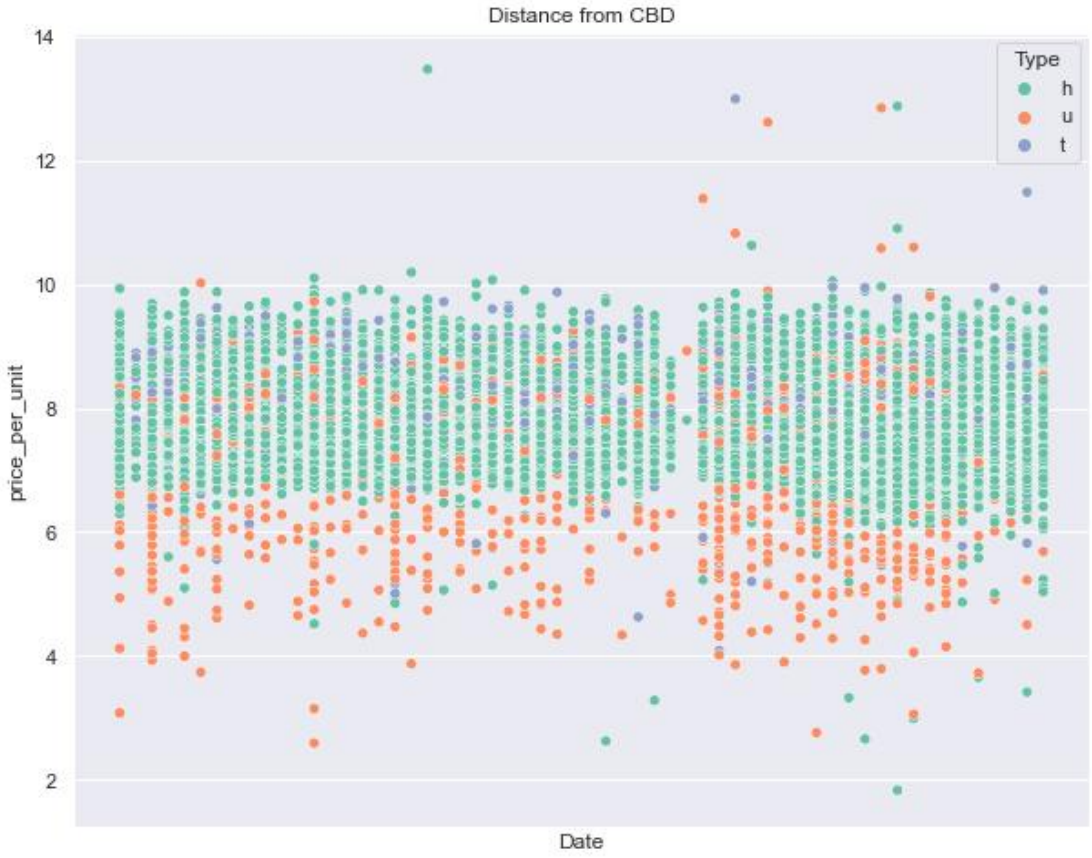
ویلاها کمتر است . همچنین در اواخر سال ۲۰۱۷ سقوط بیشتر این فروش ها را میبینیم. به همان میزان که از فروش این خانه ها کاسته شده است به فروش خانه های ویلایی افزوده شده است.

```
In [194]: df['price_per_unit']=df['Price']/df['Landsize']

In [199]: import numpy as np
          plt.figure(figsize=(10,8))
          sns.scatterplot(y=np.log(df['price_per_unit'])
                          ,x=df['Distance'],palette='tab10', data=df,hue='Regionname').set_title("Distance from CBD");
```

در اینجا به کمک ستون های قیمت و متراژ ابتدا قیمت هر مترمربع را پیدا میکنیم.

سپس نسبت قیمت به فاصله از مرکز شهر را بر روی نمودار میبریم. نقطه های رنگی انواع مکان املاک را نشان میدهد. در اینجا همانطور که در فاصله کمتر از مرکز شهر قیمت های متغیری را میبینیم، رنگ های سبز و آبی بیشتری را نیز میبینیم. میتوانیم نتیجه بگیریم که قیمت ملک در مناطق سبز و آبی (جنوب و شمال) شهر و سپس مناطق نارنجی(غرب شهر) بسیار بیشتر است و نقاط شرقی مناطق دورتری از فشردگی املاک میباشند.



در این نمودار قیمت هر متر مربع نسبت به گذر زمان بررسی شده است. همانطور که مشاهده میشود قیمت به طور کلی برای هر کدام از انواع ملک نسبتا ثابت و رو به افزایش است اما به طور کلی قیمت املاک واحدی و یونیت و سوئیت شامل موارد پایین تری میباشد.

نتیجه گیری

املاک ملبورن به طور کلی دارای مشخصات پیوسته ای میباشند. در این شهر فراوانی املاک با اتاق های ۲-۳ تایی بیشتر است و با دور شدن از مرکز شهر قیمت خانه ها افت میکند. برای خرید خانه ای با قیمت پایین تر شما میتوانید به فاصله های دورتری از مرکز شهر و به سمت انتخاب خانه های تک واحده، سوئیت و یونیت بروید. با وجود اینکه قیمت خانه ها به نسبت ویژگی هایشان پیوسته میباشد اما حتما قیمت چند خانه را بررسی کنید. احتمال اندکی برای وجود قیمت های پرت رو به پایین یا بالا وجود دارد.

منابع و مراجع:

- دیتاست https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot
- https://www.kaggle.com/code/saniaks/melbourne-house-price-eda
- https://www.datasciencemadesimple.com/create-frequency-table-of-column-in-pandas-python-2/

پایان